Escaping the Big Brother: an empirical study on factors influencing identification and information leakage on the Web

Escaping the Big Brother: an empirical study on factors influencing identification and information leakage on the Web

Francesca Carmagnola
Dept. of Computer Science, University of Turin

Francesco Osborne
Dept. of Computer Science, University of Turin and Knowledge Media Institute, The Open University

Ilaria Torre
Dept. of Computer Science, Bioengineering, Robotics and Systems Engineering, University of Genoa

Abstract
This paper presents a study on factors that may increase the risks of personal information leakage, due to the possibility of connecting user profiles that are not explicitly linked together. First, we introduce a technique for user identification based on cross-site checking and linking of user attributes. Then, we describe the experimental evaluation of the identification technique both on a real setting and on an online sample, showing its accuracy to discover unknown personal data. Finally, we combine the results on the accuracy of identification with the results of a questionnaire completed by the same subjects who performed the test on the real setting. The aim of the study was to discover possible factors that make users vulnerable to this kind of techniques. We found out that the number of social networks used, their features and especially the amount of profiles abandoned and forgotten by the user are factors that increase the likelihood of identification and the privacy risks.

Keywords
User identification; cross-site user profiling; social networks; privacy; inference of user attributes

1. Introduction

In a couple of years, online social networks (OSNs) have become a mass phenomenon [1] and the research community has increased its interests in developing methods and techniques to acquire, analyse and aggregate user attributes, with the aim to provide personalized services or to support users in various tasks. Examples are the suggestion of people with similar interests, the recommendation of resources [2] and the support to content annotation by tag suggestion [3]. Personal data are also used for security-related services, such as automatic filtering of unintended friends [4] and trust predictions in social networks [5]. However, technologies for automating the acquisition of user data can be a two-edged sword. On one hand, they can be very useful for the personalization and automation of certain tasks, as seen above. On the other hand, especially if misguided and associated with hacking techniques, they can become a threat for the user security. This dichotomy is typical of almost any new technology and often the most effective way to deal with it is to try to push the research further, to better understand the possibilities and the dangers that derive from a particular step of scientific progress. This is the scope of this paper.

Our attention concerns in particular a range of techniques which are able to support the user identification in OSNs by cross-checking her data from various sources. This approach has grown in the last few years [6, 7, 8, 9, 10], when most users started to have accounts on different OSNs. What people typically ignore is that putting together data gathered from different sources makes possible to obtain a very precise picture of their personal information, preferences and tendencies.

This raises some important questions. How effective are the identification techniques based on cross-site checking today? How can users defend their privacy and what are the main risk factors? Does the average user really care? If they do, how can they defend themselves? And if they do not, is it possible to raise the awareness by showing them the potential possibilities and risks?

In this paper we present a study that we conducted on a sample of subjects who were asked to perform a search by using a search engine prototype that we developed based on cross-site checking techniques. Afterwards subjects were required to complete a questionnaire on their experience. The results of this study are compared and integrated with the results obtained by using a larger set of data automatically extracted from an identity aggregator. The analysis of these two evaluations shows that the identification approach is very effective. Moreover, the statistical study found significant correlations between certain behaviours of users on the Social Web and the precision and recall of identification we obtained by using our cross-site checking technique. We will analyse the factors that make a user more or less vulnerable, concerning the number and the features of the OSNs used. Moreover, we will discuss how certain trends, such as trying new OSNs and then abandoning them without deleting the account, can have dramatic effects on the security of personal data. Finally, we will show that the user awareness to privacy problem can be actually increased by making people try search engines like the prototype we developed.

The paper is organized as follows. Section 2 presents the background on user identification across OSNs, section 3 describes our technique of cross-site checking and linking of profiles, section 4 evaluates the technique by showing precision and recall of the retrieved data and section 5 examines the information collected with the interviews and identifies factors that are correlated with the probability of identification. Section 6 draws the conclusions.

2. Background

According to a widely cited definition, personally identifiable information is any information which can be used to distinguish or trace an individual's identity alone or through the combination with other identifying information which is linked or linkable to a specific individual, such as date and place of birth[1]. Cross-site user identification fits the second part of the definition, since it deals with the possibility of identifying one individual by linking and combining information from different sites. It is a subject of interest for different communities, which often use similar techniques. It is relevant for user profiling in the community of adaptive and personalized systems [2, 3], in that one of criminal identification [8] and furthermore in that one of privacy protection [6, 11, 12].

Research studies on security and privacy protection analyse risk factors that make easier the unintended personal-information leakage. Zheleva and Getoor [6] discuss the "illusion of privacy in social networks", stating that while a person's profile may remain private, information and relationships that are public can leak a large amount of information. Moreover, disparate pieces of information about a person can be connected to obtain a more complete profile of that person.

The easiest way for a crawler to connect two or more profiles on different OSNs is following explicit links between profiles. This possibility is frequently used, given that profile templates on OSNs have often fields where the user can enter the URL of further profiles of her on other OSNs. Moreover, formalisms such as Friends Network (XFN) and Friend Of A Friend (FOAF), define standard structures for representing the user accounts. This makes the extraction and mapping of such data easier. The Google Social Graph APIs, used for example in [2, 13], exploit these formalisms to map users on different social sites. However, actually they are not widely adopted by OSNs and this is probably the reason for the company to abandon the Social Graph project.

When profiles are not connected explicitly, there are still chances to identify users on OSNs. Several techniques have been tried, many of them borrowed from entity resolution (ER). ER is the process of determining whether two references (e.g., two user profiles) are referring to the same entity (e.g., the same person) or not [14]. Entities are described in terms of their characteristics, called attributes, thus a reference is a collection of attribute values for a specific entity. The issue of entity resolution has been addressed for more than five decades, referred to with different names, by different communities: record linkage, record linking, record matching, deduplication, entity matching, entity linking (EL) and identity matching [15]. It is addressed in fields such as statistics, database, data integration, natural language understanding, AI, Semantic Web, etc.

ER algorithms and strategies can be classified according to different criteria: ER approach, data type used for ER and ER technique.

2.1. ER approaches

The traditional approach to resolve if two references refer to the same entity is *direct matching*, by computing pairwise similarity over the attributes of references [16, 14]. Several similarity measures have been developed for string matching, such as Levenstein distance [17] and Jaro distance [18], and for approximate string matching, such as [19]. Some of them are adaptations of other metrics suited for name and token matching (e.g., Jaro-Winkler distance [20], atomic string distance [21]). The similarity of each attribute can be computed by using the measure that works best for it

and the final similarity between references can be then computed as a weighted combination of the similarities over each attribute.

A different approach for entity resolution takes into account *relational data* concerning a reference. This approach is usually named relational ER. Bhattacharya and Getoor [16] distinguish two kinds of relational ER. They call *naïve relational ER* the approach that treats related references as additional attributes for matching (e.g., comparing the coauthor names of two references) and *collective relational ER* the approach that resolves related references jointly. The peculiarity of the latter approach is that resolution decisions affect other resolutions. A way to implement it is facing entity resolution as a clustering problem with the goal of assigning the same cluster only to references that correspond to the same entity [16]. Since the references in each cluster are connected to other references, relational similarity may also consider the clusters of all these connected references. Dong et al. [22] resolve entities by propagating relational reference-similarity decisions in a dependency graph whose nodes represent similarities between pairs of references, and whose edges represent the dependencies between the reconciliation decisions. Other approaches are variations of those mentioned above. For example ER can be obtained by linking references based on a chain of directly matched reference pairs (transitive equivalence [14]), or by combining direct matching and relational approaches [e.g., 8, 16, 23].

2.2. Type of data used for ER

The distinction between direct matching-based approaches and relational approaches is often reflected in the type of attribute, *personal* vs *social*, used for linking references. Personal data, such as demographic data, have been used for long in pairwise attribute matching, even though they have been also used in relational ER such as in [23]. Additional data sources for linking references have a social nature. Social data are embedded in people's social behaviours and include the interactions of group members [8]. Examples are co-authorship, friendship, fellowship, social tagging activity, wiki activity, etc. Recent studies in OSNs use this kind of data for identity and attribute disclosure. They usually represent OSNs as graphs with nodes (the users) and edges (the links between users). Notice that in this context, "link" stands for connection between different entities, while in EL (and so in our work) "link" stands for connection between two references that refer to the same entity. Social networks and communities are environments that can provide several data, such as labels of vertices and edges, vertex properties and attributes, neighborhood graphs, induced sub graphs and their combinations [24]. This richness of data makes possible different kinds of analyses and makes easier attribute disclosure and information leakage. Approaches and methods for identity and attribute disclosure that use social links can be various and often use a combination of methods and of attribute features. Some studies identify individuals calculating the overlap of their subgraph structure starting from the assumption that comparing the social graph structure of a node in a graph against nodes in other graphs can work as a de-anonymization tool across platforms [25, 26, 27, 11, 28]. Others combine the sub-graph structure with information about nodes [29, 6, 16, 30, 31]. In [31] user profile matching is based on conditional random fields that extensively combine profile attributes and social linkage. According to the authors, this approach performs well when profile data is poor or incomplete. Bilgic et al. [30], in the domain of scientific publications, exploit the co-authorship relation to detect duplicates and perform user identification. Zheleva and Getoor [6] analyze various techniques for attribute disclosure, showing that group membership-based approaches perform better than link-based ones. Mislove et al. [32] exploit the graph structure to identify dense communities and infer user attributes, moving from the observation that users with common attributes are more likely to be friends and often form dense communities. Li et al. [8] adopt a probabilistic relational model to extract social identity features for identity matching. The social activity and relation features are used to revise matching decisions based on personal features.

2.3. ER techniques

In the paper mentioned above, Li [8] reviews the main approaches in identity matching. First he distinguishes identity matching from entity resolution/deduplication, observing that they have different goals, which may cause different management of false positives and false negatives and different management of uncertainty when few data are available. Then he proposes a classification based on the type of attribute (personal, social) and on the technique used for identification (heuristic, distance-based and probabilistic). His work integrates a survey of Elmagarmid et al. [15] who divides matching techniques in two main categories: approaches that rely on training data to "learn" how to match the records and approaches that rely on domain knowledge or on generic distance metrics to match records. Heuristic rules that are defined by exploiting domain experts' knowledge belong to this category [8]. For example, given a set of matching conditions on attributes, if two records satisfy them with a score above a threshold, they are resolved as referring to the same entity. Heuristics can be generated also from training data, if available [15] or learned empirically.

2.4. Our approach

The technique we present for cross-site identification uses a combination of approaches. Considering the classification criteria presented in this overview, it can be classified as follows. (i) ER approach: the first step of the algorithm uses *direct-matching* of pairwise attributes and computes a weighted combination of their similarity scores; the second step of the algorithm uses a *relational approach*, by cross-linking profiles, clustering them and revising their score based on the relationship with the profiles in the same cluster. We use the term *link* as in EL: linked profiles are profiles with a certain probability of referring to the same entity. (ii) Type of attribute: the algorithm uses *personal data* (similarly to people search engines) to compare user profiles. In detail, it uses: nickname/full name, age, date of birth, place of birth and country. (iii) Technique for identity matching: the first step of the algorithm uses a *distance-based technique* to compute the similarity of nicknames and of full name, moreover it weights the match among the other attributes by computing contextual measures (persistency of the attribute and frequency of its values) that will be explained in the next section; the second step of the algorithm uses *heuristic techniques* to linking profiles, clustering them, revising their score and extracting attributes from the clusters.

2.5. Similar works

Related works that use personal attributes likewise us are [7, 9, 10, 33, 34, 35, 36]. Similarly to our first part of the algorithm, Vosecky et al. [9] use a distance-based method to calculate the similarity between profiles based on the weighted match between attributes. They represent each profile as a vector of attributes, use a set of string matching functions to calculate a similarity score between corresponding vector fields and finally assign each attribute a weight, optimized on a training dataset, to calculate the final similarity score. Differently, we weight attributes by calculating a set of indicators that will be described in the next section: attribute's persistency and attribute's value frequency.

Attributes like username and nickname, that we exploit, have been analyzed in some studies [10, 36]. The peculiarity of these attributes is their close relationship with the owner, which, in some cases, makes them unique. Username uniqueness has been investigated from Perito et al. [36], who suggest that usernames are unique enough to identify profiles across networks. To demonstrate that, they train a machine learning classifier using Markov Chains and TF*IDF and they show that the first performs better than the latter. In our approach, the uniqueness is defined in the context of a specific population instead of as an absolute property [37]. On this basis, we developed a measure that computes the nickname's specificity in a population and we use it as an indicator of the likelihood that two profiles with the same or a similar nickname belong to the same user in that population.

Iofciu et al. [35] identify users across online tagging systems by exploiting personal data (username) and social data (the set of tag assignments performed by the user, that the authors define implicit feedback). They compute the similarity on both the types of attribute and then combine them through a weighted sum. Wang et al. [33] use a record comparison algorithm for detecting deceptive identities by comparing four personal features (name, date of birth, social security number, and address) and combine them into an overall similarity score. Shen et al. [34] match user profiles (that they call "mentions") by using the values of their attributes and adopt a probabilistic approach to entity matching that exploits some domain constraints, in the form of heuristic rules, which can be either learned from the data or specified by a domain expert or user. Motoyama and Varghese [7] parse profiles on Facebook[2] and MySpace[3] and extract a set of basic attributes. However, differently from us, they train a classifier using "boosting methods" to identify profiles that may belong to the same user.

Studies on ER and identity matching are complemented by studies on risks related to user de-anonymization and attribute disclosure, which are closely related to our study with real users. Irani et al. [1] try to measure the users' average exposure to the attacks of user identification and password recovery by quantifying the possibility to retrieve in OSNs the personal information required for the attacks. The authors of the study move from the observation that many websites let users recover their passwords by providing personal information (such as birth date and address). They observe that attackers often infer the answers from other sources and use this recovery mechanism to compromise accounts. For their study, they exploit identity aggregator sites which let users manage their online identities by providing links to their various social networks. They found that the amount of information released for the physical identification attack increased from 34 percent for a user with one social network profile to 90 percent when combining six or more social profiles. As we will discuss in section 5.1 these results confirm our findings. We can also notice that an indirect result of their experiment is to show the dangerousness of making public the profiles on identity aggregator sites since it offers easy connections between the profiles of a user.

Further risks related to cross-checking techniques concern the possibility to combine personal public information acquired from OSNs with information from other sources, such as information restricted to friends or information from

other repository, e.g., institutional repository [32]. This makes possible, for example, to perform more insidious and personalized phishing attacks and malware [38].

3. Cross-site checking and linking of user profiles

This section concerns the technique we designed for user identification and retrieval of personal information on different OSNs. Given the scope of the paper, we will focus mainly on those features of the technique relating to privacy risks.

Our approach consists in cross-checking the user attributes on different profiles, retrieving profiles with compatible attributes and calculating a probability of match. The technique requires as input a variable set of attributes, such as nickname, age, gender, city, etc. of an individual to be searched. The technique returns the profiles that are likely owned by the searched user on different OSNs and a list of attributes extracted or inferred from these profiles. Both the profiles and the attributes are returned with a likelihood of match, that we call certainty factor (CF) ranging from 0 to 1.

The peculiar features of the approach are the following:

- attributes are not managed all in the same way: the match score of pairwise attributes is computed by taking into account the attribute's persistency, the frequency of its possible values and also, for nickname attributes, the attribute's specificity. For this reason, the match of different attributes may produce different match scores and moreover, the match of the same attribute on different OSNs may produce different match scores as well. Basically, the more specific, persistent and uncommon in the given context two matched attributes are, the higher the probability that the two profiles belong to the same user;
- profiles with some matching attributes, that do not have contradicting ones and share newly found attributes are clustered together and the score reassessed towards the highest score in the cluster (cross-linking of attributes of different profiles).

The features above concern respectively the first and the second macro steps of the technique, that we mentioned in the previous section. We will explain them below in more detail.

3.1. First step

A set of candidate profiles is extracted from a number of OSNs[4] and each profile is assigned a match score, based on the comparison of the profile with the input attributes. The score is computed as the linear combination of two measures: the *nickname score* and the *attribute score*.

The *nickname score* measures the similarity between the input nickname(s) and the nickname of the retrieved profile and is a function of the Levenshtein distance [17] and of the nickname specificity. The specificity is a measure derived from the length of a nickname and its degree or rarity within a specific social network or a set of them. To estimate the rarity of a class of nicknames in a context we defined different categories of nicknames composed of different patterns of letters, numbers and alphanumeric characters (see also [37, 39]). For example, in a certain OSN, a nick composed of lowercase characters and no numbers can be common, while in another OSN it can be very rare, and thus more indicative for user identification.

The *attribute score* depends on the number and quality of matching attributes. Our technique assigns a specific weight to each attribute match/mismatch, since the information derived from it varies greatly depending on the features of the attribute. For example, we intuitively know that two persons who share the hometown are more likely to be the same person that two persons who share the gender. In the same way, two persons who share a little village as hometown will be more likely the same than two people born in Los Angeles. To convert this principle into a method, we defined some indicators to compute a weight for the positive or negative match. The indicators are: the *persistency* of the compared attribute and the *frequency* of the compared attribute's value.

The *persistency* of an attribute is correlated with the probability that its value will not change over time and that it is not different in multiple profiles of the same person. Many attributes like profession or homepage can change over time and thus they are not very reliable attributes. Others, instead, should not change, unless of errors or fakes, like birth place, birth date and the national personal code. In general, a match that involves persistent attributes is more indicative for identification than a match between not persistent attributes. This is particularly true in case of negative match. A negative match that involves a user's homepage or profession can happen simply because the user changed job or homepage. On the contrary, a negative match that involves a person birth date or birth place is a strong indication that the two profiles cannot be owned by the same person.

The persistency of an attribute in a population can be computed as the frequency of variation of that attribute in different profiles of the same person. We modulate this value within a logarithmic function that expresses an inverse relation between the frequency of variations and the weight assigned to the attribute. This technique permits to find old profiles which may contain no up-to-date information, since it weights differently persistent and not persistent attributes. In section 5.3 we will see that many subjects of the experiment we carried out found profiles that they had forgotten to have.

The *frequency* of the values of an attribute is useful to estimate if a certain value is more or less common in a given population. If the frequency is low, the value is rare and thus it should receive a higher weight in case of positive match. Again we use a logarithmic function that expresses an inverse relation between the frequency of the attribute's value and the weight assigned when the match occurs.

3.2. Second step

After assigning a score to each candidate profile, the next step is cross-linking the profiles. With reference to the classification in the previous section, this phase uses a relational approach for entity resolution. It consists in cross-checking the profiles, clustering them and reassessing their scores accordingly. This step addresses the case where some of the retrieved profiles share the same value for some not input attributes. This opens to the possibility that the same person owns them. In case one of the linked profiles has a score that makes it a good candidate for user identification, then, for the transitive property, the searched user should also own the other one(s). This line of reasoning allows identifying profiles that would be impossible to consider for the identification by using simply direct matching of pairwise attributes. In fact, many of those profiles would have a low score since they contain values for just a few input attributes, but they can nevertheless contain other attributes that can be used for linking profiles.

We implement this solution by building non-exclusive clusters that contain profiles that do not contradict one another and share new attributes (namely not belonging to the input attributes). The profiles inside a cluster may increase their scores with an increment depending on attributes' weight (computed as explained above). The increment has as upper limit the difference between the maximum score in the cluster and the original score of the profile and also depends on the weight of the new attributes in common. This technique allows the identification of many profiles that would not be retrievable using a direct approach that does not cross-check data from different sources. Finally, if one or more profiles are retrieved with a sufficient confidence (CF), rules are applied to extract attributes from the clusters and computing the final CF of the retrieved profiles and of the discovered attributes.

The entire process can become recursive. The new attributes discovered after a run and having a high enough CF can be used as input attributes for a new run. This can lead to find alternative identities, including identities that the user thought not attributable to her.

Compared to the techniques for cross-site user identification described in section 3, this approach introduces some innovative elements. Besides the combination of different approaches, the main feature that distinguishes it is that metrics used for assessing the likelihood that two user profiles belong to the same person are built by using context information concerning the crawled social systems. The context of the social systems being crawled is used to compute: the frequency of attributes' values, their frequency of variation and the frequency of nickname types. This makes the metrics usable on social systems with different features (e.g., different policies for nickname composition, different cultures and name patterns, etc.).

The technique described in this section has been implemented in the CS-UDD (Cross System-User Data Discovery) prototype. We used it for the evaluation of the technique and for the study on privacy risks related to malicious use of this kind of techniques.

4. Experimental evaluation

The main objective of the evaluation was to investigate how the user behaviour, the forgotten profiles and the characteristics of different social networks reverberate on the risks for privacy and information leakage. In this section and in the following, we present two tests performed by using the CS-UDD algorithm. The first and the main one was conducted in a real setting, with subjects using CS-UDD as a people search engine. We will refer to this test as **real setting test**. Subjects had to: search for themselves by providing the CS-UDD with a standard set of personal information, mark the correct retrieved data and finally answer a questionnaire about the results they obtained and about privacy-related issues. This setup of the study allowed us to:

- evaluate the ability of the technique to identify the subjects on different OSNs and to discover information about them, by calculating the precision and recall of the retrieved data,
- investigate factors that may increase/reduce the precision and recall, or, in general terms, the probability of identification, when this kind of technique is used. This second part of the study is intended as a contribution to the literature on the issue of privacy on OSNs. The outcome of our analysis confirmed the results obtained by using similar techniques, but gave also new findings about less investigated factors, such as the importance of forgotten profiles and the change in the perception of privacy risks showed by subjects after using a search engine that highlighted such risks.

To evaluate the performance of the CS-UDD algorithm on a larger sample we also performed a second test using a dataset of online profiles obtained automatically. We will refer to this test as **identity aggregator test**, since the sample of users was obtained by parsing Profilactic[5], an identity aggregator application which enabled us to know which profiles were owned by the same person.

While the focus of the paper is on the real setting test, aimed to discover patterns and user behaviours that influence user identification, the worth of the second test is that we obtained a confirm of the accuracy of identification on a large sample, showing also that the present approach is scalable enough to be used on hundreds of thousands of profiles at once.

In section 4.1 we describe the overall design of the study, in 4.2 we describe the study samples and in 4.3 the results about user identification obtained by running CS-UDD in the two tests. Section 5 will analyse the risk factors that influence the probability of identification and will discuss privacy issues concerning the study.

4.1. Design of the experimental study

The **real setting test** has been designed to obtain two kinds of data by requiring the subjects to perform two tasks.

- Part A of the test: the subjects had to search themselves using the CS-UDD beta-testing prototype, available as a web application. Each user had to provide her name, age and city to be searched through CS-UDD on Facebook, Netlog, MySpace and Skype. The search results were then displayed and the user was asked to mark them as “right” or “wrong” and to note on which social network she really had profiles and how many. No user data were stored in the database: we just maintained the metadata about what kind of attributes and profiles are found and whether they are correct or not. So, for example, if the application correctly guessed that the user was born in Napoli and worked as a journalist according to Facebook and MySpace, we simply recorded that the attributes hometown and profession retrieved from Facebook/MySpace were correct.
- Part B of the test: each subject was asked to answer seven multiple-choice questions about her relationship with social networks and her perception of privacy issues (Table 1 in section 5).

The **identity aggregator test** was introduced to evaluate CS-UDD performance on a bigger sample. We used as starting point Profilactic, a service that allows users to collect in a dashboard the content of their profiles on several social networks. We collected Profilactic profiles that were linked both to MySpace and Flickr and we excluded profiles that did not have the two basic attributes age and location. This method made possible using a dataset larger than that one used in the real setting test, but obviously it could not provide information about the user behaviour and the user perception about privacy. Thus, the two types of evaluations are complementary.

4.2. Study sample

4.2.1. Real setting test

The sample used in the real setting test includes 101 subjects, 65 females and 36 males, ranging from 19 to 35 years old, recruited according to availability sampling strategy. Subjects are students of the School of Literature and Philosophy at the University of Turin attending the classes of basic Computer Science and HTML laboratory. Given their kind of study and their age we assume that they all have a medium level of experience with digital technologies and social networks. The total number of profiles examined by CS-UDD when searching for online profiles with the attributes of the subjects was about 20 000.

By analyzing the answers about the OSNs used by the subjects, we can outline some traits of our sample. As Figure 1 shows, the most popular social network in the sample is Facebook, with 80% of students, followed by Skype. Few users in the sample use MySpace and Netlog, but as we will see, including these slightly outdated OSNs was very useful to

discover one of the major risk factors that we found: the presence of unused and forgotten profiles. About 10% of the subjects do not use any social network.

As Figure 2 shows, 54% of the subjects use two or more OSNs at once. The average number of OSNs for user is 1.6 with a standard deviation of 0.8. Multiple profiles on the same social network are not so rare. We registered an average of 1.15 profiles for Facebook users, 1.2 for Myspace users, 1.6 for Skype users and none for Netlog users. It is worth noting that both Facebook and Skype, which supposedly use real names, have nevertheless a relatively high percentage of multiple profiles.

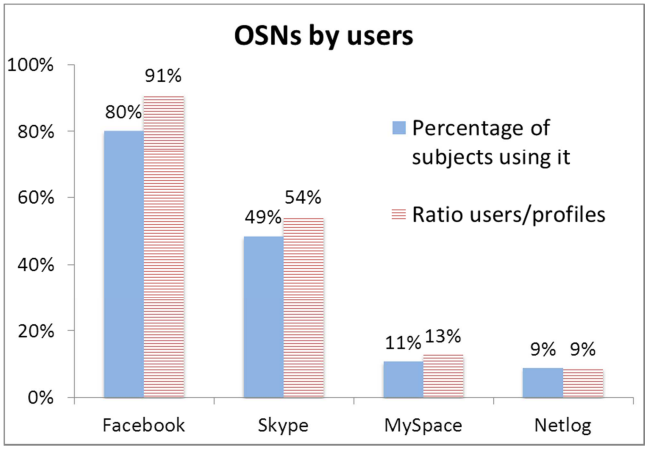

Figure 1. Distribution of users by social network.

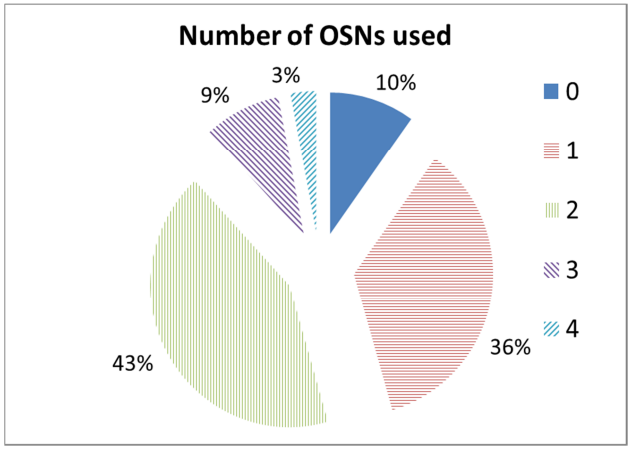

Figure 2. Number of social networks used.

4.2.2. Identity aggregator test

The sample we used for the identity aggregator test was extracted from Profilactic. The condition for a Profilactic profile to be extracted and included in the sample was to have the fields name, age, city and country filled in and to have explicit links to profiles on both Flickr and MySpace. The final sample was composed of a total of 1200 matching profiles (600 Profilactic-MySpace pairs and 600 Profilactic-Flickr pairs). CS-UDD parsed more than 120 000 profiles on MySpace and Flickr when it was run to identify the Profilactic sample users on such OSNs. We will show the results and compare them with those we obtained in the real setting test.

4.3. Results of cross-site checking and linking of user profiles

To test the ability of the technique to identify the subjects and their attributes, in both the tests we used two standard metrics in Information Retrieval: precision and recall.
The metrics are defined as:

Precision = *true positives attributes or profiles* / *retrieved attributes or profiles*
Recall = *true positives attributes or profiles* / *total true attributes or profiles*

4.3.1. Real setting test

In the real setting test, we asked the subjects to assess the correctness of the profiles and attributes returned by CS-UDD. Since each subject was asked to specify the number of profiles she has on a social network, we were able to compute both precision and recall for the returned profiles. Differently, concerning attributes, we could compute only the precision and not the recall, since most users do not know the total number of their public attributes.

The results returned by CS-UDD have a certainty factor (CF) ranging from 0 to 1, expressing the confidence of the system that the specific attribute or profile retrieved is true (CF is a support for a user or an application to make decisions about the results to trust, for example only results with a CF higher than a threshold).

Figure 3 and 4 show the results of the identification in terms of precision and recall as a function of CF. We obtain a precision ranging from 94% to 97.1% in identifying user attributes. The precision of the retrieved attributes, displayed in

Figure 3, is very high even when the algorithm had to infer them using a set of profiles with low precision, as in the case with CF threshold equal to 0.3. This result confirms that the technique to infer personal data by checking them across different probable profiles is able to reduce the false positive rate even in noisy places like social networks.

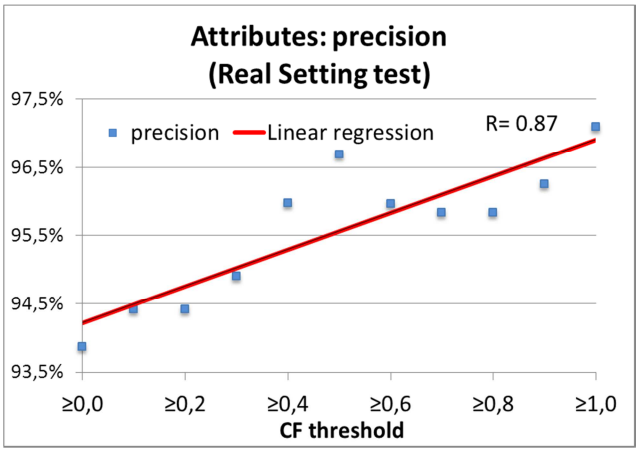

Figure 3. Precision of the returned attributes for the Real Setting test.

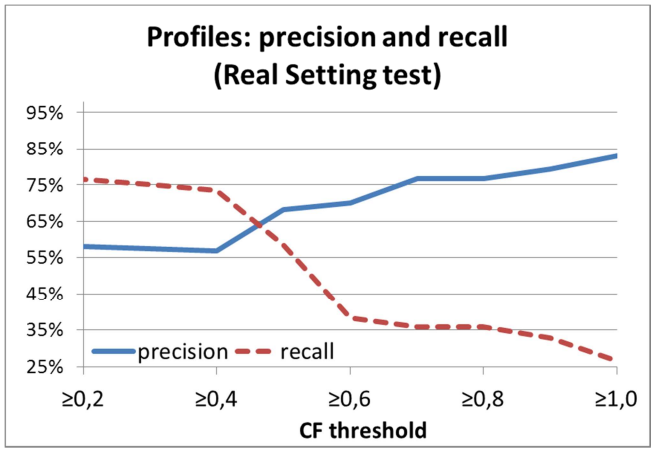

Figure 4. Precision and recall of the returned profiles for the Real Setting test.

Figure 4 displays the precision and recall of the retrieved profiles. As the graph shows, the profiles identified with maximum CF (CF=1) yield a precision of 83% and a recall of 27%. On the contrary, considering lower thresholds, the recall becomes higher while the precision decreases. However, the precision remains higher than 55% even when taking in consideration also the profiles returned with a low CF (like CF≥0.3).

A system that would implement this technique could therefore choose to trust/accept/use profiles with high CF, obtaining high precision but risking missing some profiles or, on the contrary, it could choose lower CF to enlarge the number of profiles correctly identified but with the risk of false positives. A recommender system would probably prefer a CF threshold shifted toward high values, while a human, e.g., an employer who searches for information on an employee, would probably prefer a lower CF threshold since she can check the correctness of the returned profiles by examining them personally on the Web.

4.3.2. Identity aggregator test

We can compare these results with those obtained in the identity aggregator test. Figure 5 confirms that the precision of the retrieved attributes grows linearly with the CF and it is ≥ 96% when the CF is ≥ 0.8. The average deviation of the precision yielded by the *real setting test* from the precision yielded by the *identity aggregator test* is 1.24% and its mean quadratic variation is only 1.5%. The *identity aggregator test*, however, was able to get a slightly better performance regarding the retrieved profiles. This may be due to the fact that people using online aggregators tend to take care of their different profiles and update them regularly. By comparing the recall we can see that the results relative to the smaller sample show the typical random fluctuations around the results of the larger one. However, the average deviation of the recall computed for the smaller samples from the recall computed for the larger sample is only 0.71%, while the mean quadratic deviation is 9.8%

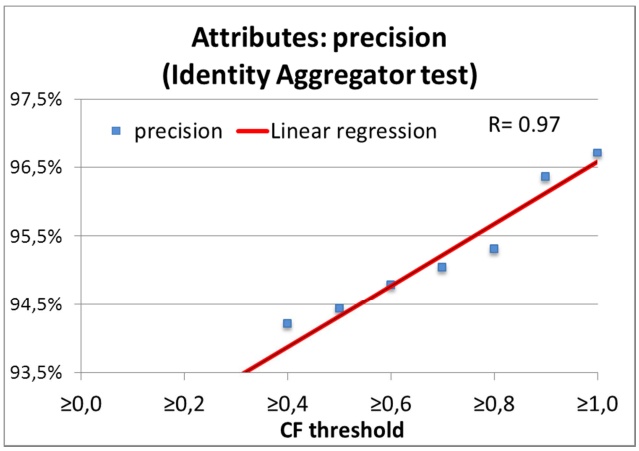

Figure 5. Precision of the returned attributes for the Identity aggregator test.

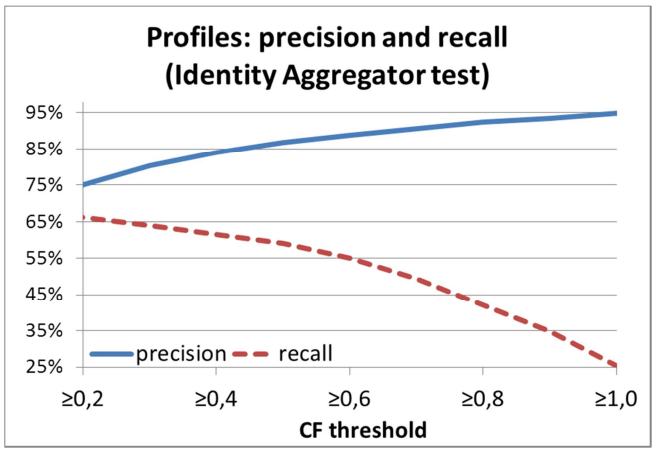

Figure 6. Precision and recall of the returned profiles for the Identity aggregator test.

Thus, we can say that the efficacy of identification, tested on the large sample of the **identity aggregator test** confirms the results obtained in the **real setting test** and in previous evaluations we carried out using preliminary versions of the algorithm [37, 39].

In the rest of the paper we will analyze **part B** of the **real setting test**, based on questionnaire, which made possible to examine user behavior and privacy risk perception.

5. User behaviour: factors that influence the likelihood of identification

In this section we present the questionnaire that subjects were required to complete in the real setting test after using CS-UDD. The questionnaire was aimed to: 1) study the relations between the user behaviour on OSNs and the probability of identification, in order to discover (risk) factors that may increase it 2) investigate the perception of users about the privacy problem related to the spreading of their public data and how it could change.

Table 1 shows the questionnaire. While the first three questions were asked to the full sample of 101 people, the remaining questions were necessarily limited to people (91 users) who actually declared to use at least one social network.

Table 1. Questionnaire to investigate the probability of identification in relation to user behaviours on OSNs.

1. In your opinion, how severe are the risks for privacy when using a social network?	Very risky: 38.6%; Risky: 52.5%; Not very risky: 7.9%; No risk at all: 1%
2. Have you ever used a people search engine (such as 123people)?	Yes: 19.8%; No: 80.2%
3. Have you ever used a profile aggregator, that is a service which merges your profiles from multiple social networks?	Yes: 12.9%; No: 87.1%
4. Did you find any profiles you forgot to have?	Yes: 20.9%; No: 79.1%
5. Did you find any personal data you thought not to be retrievable?	All: 17.6%, Some: 3.3%, At least one: 9.9%, No one: 69,2%
6. Before the test, did you know that the retrieved data were public or did you think they were private (namely visible only to your contacts)?	All Private: 12.9%, Both: 70.3%, All public: 16.8%
7. Are you going to edit your public data after seeing CS-UDD search results?	Yes: 25.3%, No: 74.7%

In order to discover possible factors that increment the likelihood of identification and of information leakage, we combined the data we obtained in part a) and part b) of the test. In detail we used the following *SOURCE DATA*:

- source data [a]: precision and recall results (section 4.3),
- source data [b]: data about the profiles owned by the subjects (sections 4.1 and 4.2),
- source data [c]: answers collected with the questionnaire in Table 1 (section 5).

This analysis enabled us to study and discover possible factors that influence privacy issues as the probability of being identified and the user perception of privacy risks. We thus could derive many significant relations between the behaviour of users in OSNs and the privacy risks. Table 2 shows these relations expressed as factors, effects and statistical significance. The table also shows some effects that resulted not significant in this particular test, but which appear to point to a real phenomenon that may be confirmed by further experiments. The first five rows refer to the main risk factors from the perspective of information leakage, while the latter two regard the user reactions to the privacy issue.

Table 2. Factors and effects derived from the combined analysis of data. In the fourth column, bold is used to indicate statistical significance according to the chi-square test ($p<0.05$).

Factors	Effects	Source data	Null hypothesis probability
Number of OSNs used by the subjects	Directly correlated to the identification risk (precision) ($r = 0.94$)	[a] & [b] (Section 5.1, Figure 8)	**p=0.03**
Dimension of the OSNs used by the subjects	Inversely correlated to the identification risk (precision) ($r = -0.98$)	[a] & [b] (Section 5.2, Figure 9)	p=0.62
Presence of forgotten profiles	Higher identification risk (recall/precision)	[a] & [c] (Section 5.3, Figure 10)	**p=0.03**/0.21
Presence of forgotten profiles	Higher risk of private attribute disclosure	[c] (Section 5.3)	**$p = 9*10^{-5}$**
Paradox of the informed users	Slightly higher identification risk (precision)	[a] & [c] (Section 5.3)	p=0.82
Unexpected retrieval of personal data	Help the awareness of privacy problem	[c] (Section 5.4, Figure 11)	**p = 0.007**
Unexpected retrieval of personal data	Intention to fix privacy problem	[c] (Section 5.4, Figure 12)	**$p = 10^{-4}$**

In the following we will describe these factors and effects. We remind that other risk factors have been studied in the literature, such as the disclosure of friendship links and of group membership, mentioned in section 2.

5.1. Number of social networks used

The central idea of the technique presented in this paper is the cross-check of different sources for discovering user profiles and attributes. The hypothesis is that the probability of identification becomes higher if a user has different profiles on different social networks. Under the perspective of user privacy, this hypothesis would mean that using a high number of social networks increases the user vulnerability.

An obvious risk is related to the simple fact that it is more likely to find at least one profile of a person if she has profiles on several different OSNs. However the number of OSNs used affects also the precision of the identification, which is increased by linking the profiles discovered on different OSNs and clustering them, as described in section 3.

Figure 7 provides an overview of the likelihood to identify birth date, university and hometown in the real setting test, considering all the OSNs, while Figure 8 displays the precision of identification as a function of the number of OSNs used, showing that our hypothesis is true. The precision of both attributes and profiles increases with the number of social networks. In particular, when the number of OSNs used is at least two, we obtain a significant boost in the precision of the retrieved profiles. This is an important result since the average number of OSNs per person is, in our sample, 1.6, with 42% of subjects having accounts on 2 OSNs and 18.8% on 3 or 4.

This result confirms the study of Irani et al. [1] on sets of profiles obtained by crawling identity management sites which let users manage their online identities by providing links to their various social networks.

Considering Figure 7, it should be noted that the percentage reported is not the recall, but an underestimation of it. In fact we do not know the number of users who had that particular attribute somewhere, and thus we computed simply the

ratio between the number of users for which the attribute was correctly retrieved and the total number of users. Nevertheless, CS-UDD was able to retrieve university and hometown for about one quarter of the users who had at least one public attribute and the birth date for 12% of them. As expected, there is an appreciable increment in the probability of finding one of these attributes when considering users who have at least one public attribute.

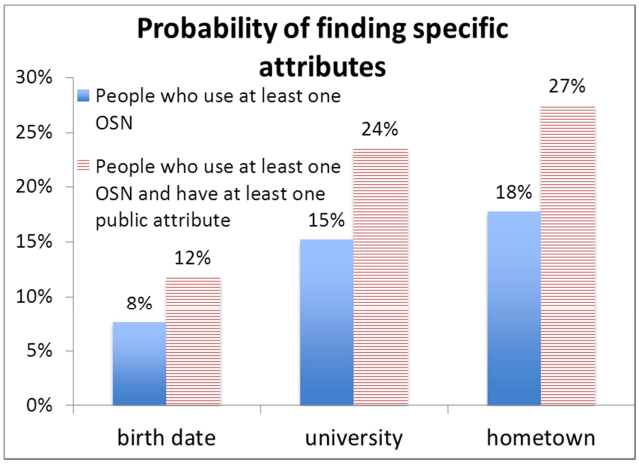

Figure 7. Probability of finding specific attributes.

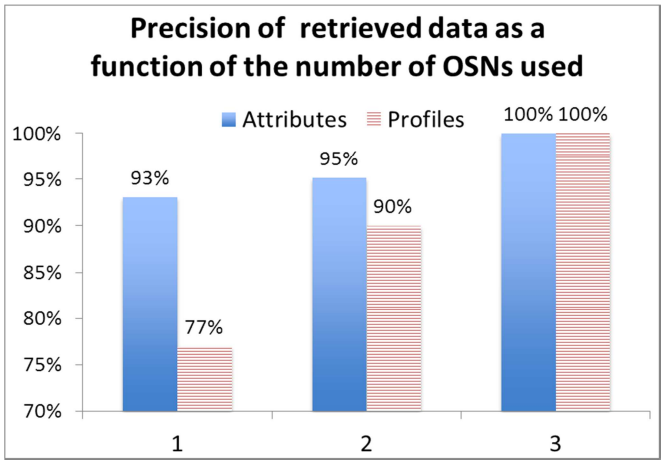

Figure 8. The precision as a function of the n. of OSNs used.

5.2. Social networks characteristics

Figure 9 shows the recall and precision for the user profiles split by social network. The precision appears to be inversely correlated with the estimated number of users for each social network (Pearson linear correlation coefficient r=-0.98). In fact we have the lowest precision for Facebook (908 million users at the moment of the experiment[6]) and then in ascending order for Skype (663), Netlog (95) and Myspace (30).

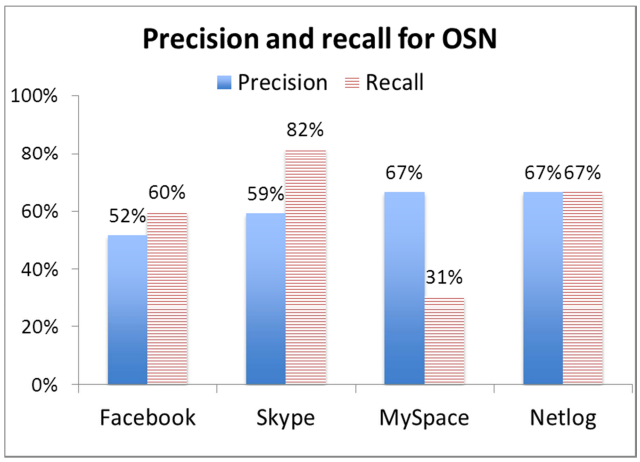

Figure 9. Recall and precision for social network.

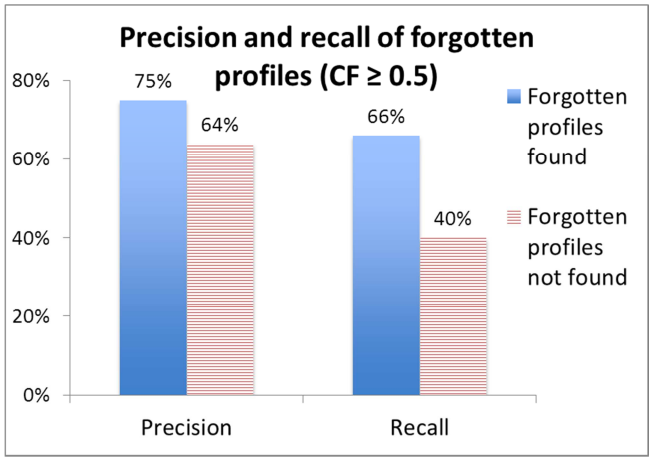

Figure 10. Relation between forgotten profiles and precision and recall of retrieved profiles.

Thus it results that a user profile is more protected in a huge social network, since the presence of homonyms and users with similar attributes makes the identification process harder. The recall moves in a different way and it is probably associated to the default data that an OSN shows and how often the attributes are updated. Skype seems particular vulnerable, since it offers by default the real name and city of a user. On the contrary, MySpace users are

more prone to use nicknames and often show very few data, making harder the retrieval of a correct user profile starting from the real name.

5.3. The ghost towns of forgotten profiles: a new threat to privacy

A significant result of our study is the discovery of the relevant role of forgotten profiles, i.e. those profiles scattered on OSNs that the user does not remember to have. This is a relative new phenomenon. In fact in the last few years many popular OSNs started to be less popular and lost active users. People typically follow such trends. They create a profile on the current top social network and then leave it when it is no longer popular. This gives rise to ghost towns of neglected profiles that can be a severe threat to privacy. Often these profiles contain data that a user may not want any more to be public or information that could be used to infer them. Besides, overt time, privacy politics of a social network can change and unfixed bugs can be exploited to obtain user data.

The presence of out of date information could suggest that this kind of profiles is misleading and thus may reduce the probability of identification. However, since the technique we presented takes into consideration the difference between persistent and not-persistent attributes, it is able to downplay the outdated information problem, performing well also on this class of profiles.

More than 20% of the subjects found at least one forgotten profile (see Table 1, question 4). As Figure 10 shows, these subjects are characterized by a precision and recall respectively 11% ($p=0.21$) and 26% ($p=0.03$) higher than those who did not find them. The reason why forgotten profiles are easier to be found may depend on the fact that old OSNs were less concerned with privacy policies, also the user might have been less concerned about the privacy problem at the time s(he) created the account and did not set personal data as private.

Forgotten profiles not only make the identification easier, but may expose the most critical kind of data: the one that the user does not even realize to be public. Answers to question 5 of the questionnaire, about "unexpected personal data found by the subjects", returned that 17,6% of subjects thought that all the retrieved attributes about her were private or, anyway, not retrievable.

Specifically, we found out that people who had at least one forgotten profile were also much more likely to find "unexpected personal data" (namely personal data they thought not to be retrievable online). To analyse this phenomenon, we split the results of the test in two groups: results from subjects who found at least one forgotten profile and results of subjects who did not. The first group was surprised by the retrieved data in 74% of the cases, while the second group only in 22% of the cases. The relative data arranged in a 2x2 contingency table (no data found vs at least one found) and analysed by the chi-square test (with Yates correction) yields a statistically significant difference ($p = 0.00009$) and an associated risk ratio $RR = 3.4$ (95%CI 2.0÷5.8 not including 1). Thus, people who have one or more forgotten profiles have a probability 240% higher of finding data they thought not to be retrievable than people who did not find any forgotten profile. This proves that it is important to delete the profiles that are no longer being used since they can become a great risk for privacy.

5.4. The perception of privacy risks does not make immune

In this section we deepen the issue of privacy perception and examine the relation with the probability of user identification. Looking at Table 1, the privacy issue is considered as important by more than 90% of the subjects. However, as we will see, often this perception does not imply a proactive action to fix the problem. Thus it seems that, while the user does know about the problem, she does not know how to fix it or she does not have the tools to do it.

80% of users never used a people search engine or a user profiles aggregator (an identity management system). Thus, many people who use social networks may not be aware about the tools to keep control of their data or search them. And while they may care about protecting their personal data on a single social network, they often ignore that collecting information from different sources may enable to infer many personal information.

We used questions 2 and 3 to define an informed user as a subject that used a people search engine or an aggregator. We then analysed the difference of recall and precision between informed users and the others and found that while the recall was similar, the precision was about 10% higher for the former. This may derive from the fact that the informed users have an average number of accounts on OSNs equal to 2.0 whereas other users have 1.5 and, as we explained in section 5.1, the number of social networks is an important factor of risk. Hence, as a paradox, it seems that being an informed user and knowing about tools like people search engine does not make this subject any safer.

Questions number 4, 5 and 6 address specific aspects of the problem of the traceability of personal data: often a user does not know which data she has online, which ones are public and which private. More than 30% of users in our sample found personal data they thought not to be retrievable. Question 6 shows that 80% of the subjects thought that at

least some of the retrieved personal data were private. But, as stated before, the technique enables to retrieve only public data. This means that the users are confused about which data are actually public. One of the reasons is that social network tools for setting the privacy level are often not very clear, even for users with some expertise.

Finally, we found that about one quarter of the subjects decided to fix their personal data after trying CS-UDD search engine. We consider this a good success in the process of making people more conscious about their data in OSNs. If more people could exploit modern people search engines to find out how exposed they are in the Social Web, the attention to privacy would surely increase.

5.4.1. The perceived risk: "It can't happen to me!"

Now that social networks are so widespread, even among minors, it becomes vital to teach users to recognize the status of their data and vulnerabilities. Today this issue is well known even among average users: many people read on newspaper news about employee fired for what they wrote on OSNs. However few people take actions to fix and check their data.

This study suggests that one of the main reasons is that, while the user perceives the danger, she does not feel it near to her experience. It is the psychological mechanism of "It can't happen to me", according to which people tend to underestimate the risk for themselves in many instances, e.g., in car accidents [40].

Systems like CS-UDD can help to break this illusion by showing that the user perception about what data are retrievable and what personal information can be found are wrong. Figure 11 shows how the awareness of the privacy problem changes dramatically for the user that found some unexpected personal data. No user that found this kind of data dismissed the privacy issue as not risky. The risk was considered severe by over 64% of these users, but only by 33% of users who did not had the same experience. The difference between the two groups is statistically significant, with $p = 0.007$.

The same dynamic is found even stronger for the decision to fix personal data after this experience. As Figure 12 shows, 57% of the people who found unexpected personal data decided to reassess their online data, against 11% of the people who did not find any of them ($p = 10^{-4}$).

This test was run over groups of 20-30 students at once, thus the subjects could confront their experiences and see the distress of the ones to whom the application recovered personal data. However, almost only students that found them in first person decided to act.

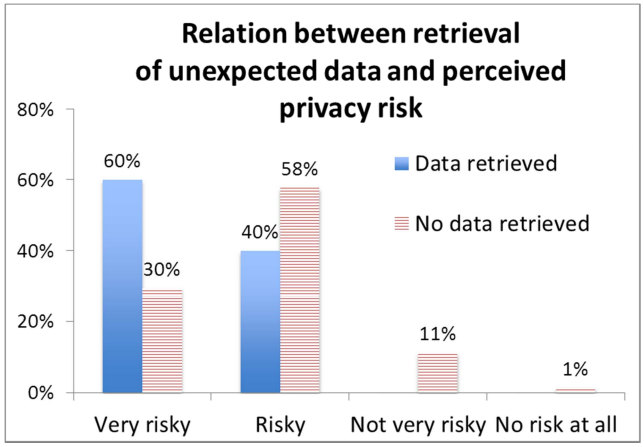

Figure 11. The shift in the perception of the risk depending on the search results.

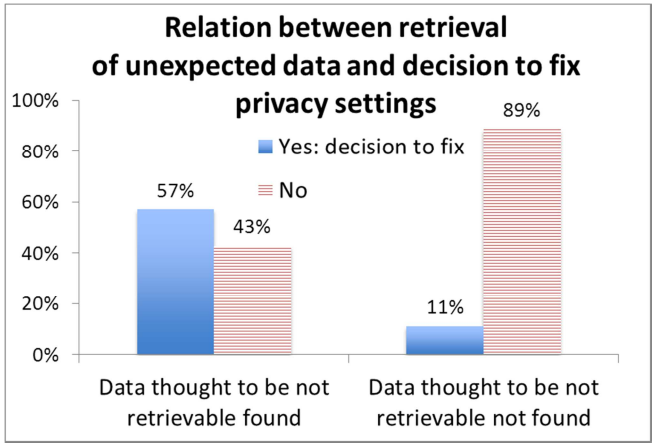

Figure 12. The effect of finding not expected personal data on the decision to fix privacy settings.

This is another example of "this can't happen to me": the fact that CS-UDD could not retrieve a particular user data does not mean that other engines could not. Moreover, the tutor reported that students who tried to improve the control over their public data had difficulties and got discouraged, since OSNs often lack proper user-friendly tools.

6. Conclusion

In the last years, OSNs have been increasingly used by people, with the result that many personal data are stored on these systems.

Even if social network's privacy settings ensure to protect the user profile information [1], they do not ensure protection from attackers who may combine disparate pieces of information about a user from multiple networks, thus allowing the user identification and the retrieval of user personal data. Cross-site user identification may be very useful for the optimization of some tasks requiring user modeling, such as user support and personalization. However, it can also be used with criminal purposes thus representing a risk for user privacy.

In this paper we presented a technique we developed to enable cross-site user identification and retrieval of unlinked personal information by connecting profiles of a user on different OSNs. The evaluation both in a real setting and by using an identity aggregator showed that the technique is very effective, resulting in a high probability of exploiting public data in OSNs to identify users and collect their personal data.

Starting from this result, we investigated the possible risks that make the user more vulnerable to cross-site user identification techniques. We found out that the risk of identification, and thus the risk of leakage of personal data, increases or decreases when varying some factors, such as the number and the features of the social network used and the presence of forgotten profiles in OSNs.

Finally, the test has also shown that the user awareness to the privacy problem can be increased by fostering users to use people search engines to monitor their own data scattered in OSNs.

Notes

1. OMB Definition (from M-07-16), http://www.whitehouse.gov/sites/default/files/omb/memoranda/fy2007/m07-16.pdf.
2. http://www.facebook.com.
3. http://www.myspace.com.
4. Profiles can be retrieved in different ways: by exploiting the search tools provided by the OSNs, by using ad hoc crawlers, etc.
5. http://www.profilactic.com/.
6. http://www.telecompaper.com/news/.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

References

[1] Irani D, Webb S and Pu C. Modeling Unintended Personal-Information Leakage from Multiple Online Social Networks. IEEE Internet Computing 2011; 15: 13-19.

[2] Abel F, Aurojo S, Gao Q and Houben G. Analyzing Cross-System User Modeling on the Social Web. In: Proc. of the 11th International Conference on Web Engineering. Paphos, Cyprus: Springer-Verlag, 2011, pp.28-43.

[3] Stewart A, Diaz-Aviles E, Nejdl W, Balby Marinho L, Nanopoulos A and Schmidt-Thieme L. Cross-tagging for personalized open social networking. In: Proc. of the 20th ACM Conference on Hypertext and Hypermedia. New York, USA: ACM, 2009, pp.271-278.

[4] Shehab M, Ko M and Touati H. Enabling cross-site interactions in social networks. Social Network Analysis and Mining 2012; 2: 1-14.

[5] Borzymek P, Sydow M and Wierzbick A. Enriching Trust Prediction Model in Social Network with User Rating Similarity. In: Proc. of the International Conference on Computational Aspects of Social Networks. Fontainebleau, France: IEEE Computer Society, 2009, pp.40-47.

[6] Zheleva E and Getoor L. To join or not to join: the illusion of privacy in social networks with mixed public and private user profiles. In: Proc. of the 18th International Conference on World wide web. New York, USA: ACM, 2009, pp.531-540.

[7] Motoyama M and Varghese G. I seek you: searching and matching individuals in social networks. In: Proc. of the 11th International Workshop on Web information and data management. Hong Kong, China: ACM, 2009, pp. 67-75.

[8] Li J, Wang G A and Chen H. Identity matching using personal and social identity features. Information Systems Frontiers 2011; 13: 101-113.

[9] Vosecky J, Hong D and Shen V. User identification across multiple social networks. In: Proc. of the 1st International Conference on Networked Digital Technologies. Ostrava, The Czech Republic: 2009, pp.360-365.

[10] Zafarani R and Liu H. Connecting corresponding identities across communities. In: Proc. of the 3rd International Conference on Weblogs and Social Media. San Jose, California: AAAI Press, 2009, pp.354-357.

[11] Narayanan A and Shmatikov V. De-anonymizing Social Networks. In: Proc. of the 30th IEEE Symposium on Security and Privacy. Oakland, California, USA: IEEE Computer Society, 2009, pp.173-187.

[12] Balduzzi M, Platzer C, Holz T, Kirda E, Balzarotti D and Kruegel C. Abusing social networks for automated user profiling. In: Proc. of the 13th Symposium on Recent Advances in Intrusion Detection. Ottawa, Ontario, Canada: Springer-Verlag, 2010, pp.422-441.

[13] Szomszor M, Alani H, Cantador I, O'Hara K and Shadbolt N. Semantic modelling of user interests based on cross-folksonomy analysis. In: Proc. of 7th International Semantic Web Conference. Karlsruhe, Germany: Springer-Verlag, 2008, pp.632–648.

[14] Talburt J. Entity resolution and information quality. 1st ed. San Francisco, CA, USA: Morgan Kaufmann, 2010, pp. 235.

[15] Elmagarmid A K, Ipeirotis P G and Verykios V S. Duplicate Record Detection: A Survey. IEEE Transactions on Knowledge and Data Engineering 2007; 19: 1-16.

[16] Bhattacharya I and Getoor L. Entity Resolution in Graphs. In: D. J. Cook and L. B. Holder (eds) Mining graph data. Hoboken, NJ, USA, 2006.

[17] Levenshtein V. Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics Doklady 1966; 10: 707-710.

[18] Jaro M A. Unimatch: A Record Linkage System: User's Manual. Technical report, Washington, D.C.: 1976, p. 275.

[19] Navarro G. A guided tour to approximate string matching. ACM Computing Surveys 1999; 33: 31–88.

[20] Winkler W E. String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage. In: Proc. of the Survey Research Methods Section. American Statistical Association, 1990, pp.354–359.

[21] Monge A E and Elkan C P. The Field Matching Problem: Algorithms and Applications. In: Proc. of the International Conference on Knowledge Discovery and Data Mining. Portland, Oregon: AAAI Press, 1996, pp.267-270.

[22] Dong X. A. Halevy and J. Madhavan, Reference reconciliation in complex information spaces. In: Proc. of the 24th ACM SIGMOD International conference on Management of data. Baltimore, Maryland, USA: ACM, 2005, pp.85-96.

[23] Ananthakrishna R, Chaudhuri S and Andganti V. Eliminating fuzzy duplicates in data warehouses. In: Proc. of the 28th International Conference on Very Large Data Bases. Hong Kong, China: VLDB Endowment, 2002, pp.586-597.

[24] Sharma S, Gupta P and Bhatnagar V. Anonymisation in social network: a literature survey and classification. International Journal of Social Network Mining 2012; 1: 51-66.

[25] Labitzke S, Taranu I and Hartenstein H. What your friends tell others about you: Low cost linkability of social network profiles. In: Proc. of the 5th International ACM Workshop on Social Network Mining and Analysis. San Diego, California, USA: 2011, pp. 51-60.

[26] Hay M, Miklau G, Jensen D and Towsley D. Resisting structural re-identification in anonymized social networks. International Journal on Very Large Data Bases 2008; 1: 102-114.

[27] Liu K and Terzi E. Towards identity anonymization on graphs. In: Proc. of the 28th International Conference on Management of data. Vancouver, Canada: ACM, 2008, pp.93- 106.

[28] Veldman I. Matching Profiles from Social Network Sites. Master's thesis. University of Twente, Nederlands, 2009, p. 128.

[29] Kozikowski P and Groh G. Inferring Profile Elements from Publicly Available Social Network Data. In: Proc. of the 3rd IEEE International Conference on Privacy, Security, Risk, and Trust. Boston, USA: IEEE, 2011, pp.876-881.

[30] Bilgic M, Licamele L, Getoor L and Shneiderman B. D-dupe: an Interactive Tool for Entity Resolution in Social Networks. In: Proc. of 13rd IEEE Symposium on Visual Analytics Science and Technology. Limerick, Ireland: Springer-Verlag, 2005, pp. 505-507.

[31] Bartunov S, Korshunov A, Park S T, Ryu Wand Lee H. Joint link-attribute user identity resolution in online social networks. In: Proc. of the 6th Workshop on Social Network Mining and Analysis. Chicago, United States, 2013.

[32] Mislove A, Viswanath B, Gummadi K P and Druschel P. You are who you know: inferring user profiles in online social networks. Proc. of the 3rd ACM International Conference on Web search and data mining. New York, USA: ACM, 2012, pp.251-260.

[33] Wang G A, Chen H, Xu J J and Atabakhsh H. Automatically detecting criminal identity deception: an adaptive detection algorithm. IEEE Transactions on Systems, Man, and Cybernetics Part A: Systems and Humans 2005; 36: 988-999.

[34] Shen W, Li X and Doan A. Constraint-based entity matching. In: Proc. of the 20th National Conference on Artificial intelligence. Pittsburgh, Pennsylvania, USA: AAAI Press, 2005, pp.862-867.

[35] Iofciu T, Fankhauser P, Abel F and Bischoff K. Identifying Users Across Social Tagging Systems. In: Proc. of the 5th International Conference on Weblogs and Social Media. Barcelona, Catalonia, Spain: AAAI Press, 2005, pp. 522-525.

[36] Perito D, Castelluccia C, Kaafar M and Manils P. How unique and traceable are usernames? In: Proc. of the 11th Privacy Enhancing Technologies Symposium. Waterloo, Canada: Springer-Verlag, 2011, pp.1-17.

[37] Carmagnola F, Osborne F and Torre I. Cross-Systems Identification of Users in the Social Web. In: Proc. of the 8th IADIS International Conference WWW/Internet. Rome, Italy, 2009, pp. 129-134.

[38] Luo W, Liu J, Liu J and Fan C. An Analysis of Security in Social Networks. In: Proc. of the 8th IEEE International Conference on Dependable, Autonomic and Secure Computing. Chengdu, China: IEEE Computer Society, 2009, pp. 648-651.

[39] Carmagnola F, Osborne F and Torre I. User data distributed on the social web: how to identify users on different social systems and collecting data about them. In: Proc. of the 1st International Workshop on Information Heterogeneity and Fusion in Recommender Systems. Hong Kong, China: ACM, 2010, pp. 9-15.

[40] Greening L and Chandler C. Why It Can't Happen to Me: The Base Rate Matters, But Overestimating Skill Leads to Underestimating Risk. Journal of Applied Social Psychology 1997; 27: 760–780.